A "Cognitive Miser" Theory of Cooperators' Advantage
Author(s): John Orbell and Robyn M. Dawes
Source: *The American Political Science Review*, Vol. 85, No. 2 (Jun., 1991), pp. 515-528
Published by: American Political Science Association
Stable URL: http://www.jstor.org/stable/1963172
Accessed: 03/08/2013 10:25

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Political Science Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Political Science Review.*

http://www.jstor.org

All use subject to JSTOR Terms and Conditions

A "COGNITIVE MISER" THEORY OF COOPERATORS' ADVANTAGE

JOHN ORBELL
University of Oregon
ROBYN M. DAWES
Carnegie Mellon University

We propose a new model of cooperators' advantage depending neither on supplementary incentives nor on cooperators' capacity to recognize, and play selectively with, other cooperators. It depends, rather, on players' making the play–no play decision by the heuristic of projecting their own "cooperate–defect" choices onto potential partners. Cooperators offer to play more often, and fellow cooperators will more often accept their offer. When certain boundary conditions are met, this results in a higher expected payoff for cooperators than for defectors. Empirical support for this heuristic is suggested by expectations data from related social dilemma experiments. Moreover, its use can be justified in Bayesean terms. Our model brings behavioral decision theory's "cognitive miser" paradigm to bear on interdisciplinary concern with the evolution of cooperative behavior and shows how, if other mechanisms provide a suitable "initial kick," cooperation can evolve in the absence of iteration and in large, mobile societies.

> Every man is in his own person the whole human race, with not a detail lacking. I am the whole human race without a detail lacking; I have studied the human race with diligence and strong interest all these years in my own person; in myself I find in big or little proportion every quality and every defect that is findable in the mass of the race. . . . What a coward every man is! and how surely he will find it out if he will just let other people alone and sit down and examine himself.
>
> —Mark Twain

For cooperation to survive and prosper, standard theory requires that it must, somehow, rebound to the comparative advantage of the cooperator. But that is, at best, problematic. In the prisoner's dilemma terms in which the problem is classically formulated, cooperation is a dominated choice; those who cooperate are worse off, by definition, than those who defect. Thus, cooperation appears to be at an evolutionary disadvantage and cooperative tendencies should vanish over time. That outcome is, of course, collectively undesirable because aggregate payoffs are highest in the cooperate–cooperate cell and lowest in the defect–defect one.

Yet cooperation *does* survive, and a major agenda of contemporary social theory has been to explain how it does. Many theorists propose that it happens because the dilemma's incentives are supplemented by others sufficient to make cooperation more rewarding *in net* than defection:

1. Trivers (1971) has shown how kin altruism can promote inclusive or genetic fitness, and Alexander (1979, 1987) has shown how, in the context of indirect reciprocity, cooperation can prove privately advantageous;
2. theories of social evolution (e.g., Boyd and Richerson 1988; Campbell 1975; Richerson and Boyd 1988; Singer

1983), as well as sociological theory more generally (e.g., Parsons 1937), propose that cooperation is maintained by sanctioning systems sufficient to make it, in net, better for the acting individual than noncooperation;

3. Taylor (1976) and Axelrod (1984) have argued—assuming "reciprocal altruism"—that iteration of the game can modify the prisoner's dilemma incentives of single-play games so that contingent cooperation is to each player's private advantage;
4. political theory recognizes that the state (Leviathan) has the capacity to convert the prisoner's dilemma incentives of the state of nature into a nondilemma in which contributing to public goods (in Hobbes's [1947] classic case, social peace) is the rational egoist's dominant choice, so that state power has an ethical mandate based on support from people who understand —explicitly or implicitly—that without it they face multiple prisoner's dilemmas in their interactions with others (see, e.g., Buchanan 1972);
5. elaborating the Hobbesian argument, modern constitutional theory has sought rules for collective decision making that provide private incentives supporting collectively advantageous outcomes ("incentive-compatible designs")—which rules might be deliberately constructed or might evolve as a result of their beneficial consequences for the population (see, e.g., Brennan and Buchanan 1985; Buchanan 1977).

If the individual *must* play the game, the egoistically rational choice is, by definition, to defect. Under that critical constraint, therefore, the *only* way by which cooperation can end up ahead is through a change of the net incentive structure—as proposed by theories such as those just outlined.

A second paradigm is based not on the choice between cooperation and defection per se but on the choice people *can* often make between playing and not playing the game. Peoples' capacity to recognize and play with cooperators and to recognize and *not* play with defectors is pivotal here, and several authors have shown how mechanisms that promote accurate expectations about others' behavior can produce a cooperators' advantage. Frank (1988), for example, has suggested that evolved emotions serve as cue givers about the probable choices of others, thus letting cooperators avoid interaction with defectors and reap the benefits of mutual cooperation. In the same vein, Gauthier (1986) has discussed *translucency*, a characteristic of behavior that provides a basis for including or excluding individuals from cooperative play. Note that this second, selective-play paradigm depends on the existence of a "cooperative surplus" (Gauthier 1986), that is, on a larger net payoff being available at mutual cooperation than at any other outcome. In terms of this paradigm, however, cooperation remains strictly a dominated choice.

We propose a new model that is cast in the broad terms of the selective-play paradigm, but in which the cooperators' advantage does not depend on cooperators' capacity to recognize other cooperators or to unmask would-be defectors. In our model, cooperators win but not because they and their targets are, to some degree, translucent. In fact, while our model requires both cooperators and defectors to *believe* they can spot cooperators, it builds from the conservative assumption that their performance is no better than random with respect to the facts of target behavior. That is, cooperators' and defectors' judgments about whether someone else is a cooperator or defector are *noncontingent.* Our selective-play model of cooperators' advantage is, therefore, distinct from Frank's (1988) and Gauthier's (1986) versions, both of which depend on expectations about successive individuals having some predictive accuracy.

Instead of translucency, our model assumes that (1) judges project their *own* choices between cooperation and defection onto the choices of targets, that is, judges' expectations are developed from the heuristic that their own choices are diagnostic (Dawes 1988a) of others' behavior; and (2) judges (cooperators and defectors alike) make their individual-by-individual play–no play decisions on that basis. In these terms, as we will show, it is possible for cooperators to have higher expectations across a set of *options* to play than do defectors despite the higher per-play payoff that, by definition, adheres to defection.

The model has several attractive features.

1. It identifies circumstances under which cooperators will end up ahead of defectors without the awkwardness of resolving the dilemma by eliminating it and without the information demands of models that depend on translucency. As we will show, the process we identify is quite compatible with such more standard processes; indeed, it can build on whatever base of cooperation those processes establish—even needs them to operate before it can operate. It is, however, in many ways a much *simpler* process.
2. It brings behavioral decision theory's "cognitive miser" (Tetlock 1979) paradigm to bear on sociobiology's concern with the survival of cooperation. During the past 20-or-so years, experimental studies have challenged subjective expected utility as a descriptive model of human behavior, identifying a set of heuristics, or rules of thumb, that can be explained (and perhaps justified) insofar as they economize on cognitive effort (for surveys, see Bell, Raiffa, and Tversky 1988; Dawes 1988a; Hogarth and Reder 1987). Until now, however, this work has been independent of the no-less-formidable work in human sociobiology, whose major problem, according to Edward Wilson (1975), is to explain cooperative behavior.
3. It demonstrates how factually wrong belief in one's own judgment can help as well as hurt, consistent with David Wilson's (1980) general insight that an unrealistic mental model may have adaptive features.
4. It shows that cooperation *can* survive and flourish even among strangers interacting in large mobile societies. The model's implications are distinct, therefore, from those of other models that find support for cooperation in small face-to-face groups and that base cooperation in larger societies on extrapolation from there (e.g., Masters 1989; see also Buchanan 1965, 1978).

The Model

Parameters

Assume a population of n people who have the capacity to play two-person prisoner's dilemma games with each other, the games having payoffs $t > c > 0 > d > s$ (with the standard meanings).[1]

ASSUMPTION 1. *Each person makes a single decision whether to cooperate or defect across the set of such games—if they play them. Cooperators always cooperate and defectors always defect.*

People might cooperate because of normative or ethical considerations or because of selective incentives provided by Leviathan or some other institution or in response to internalized altruistic motivations. And they might defect because such mechanisms are insufficient to persuade them otherwise. But reasons for their cooperate–defect decisions are exogenous to the model. We assume simply *some* incidence of cooperation and of defection.

The assumption that people make a single cooperate–defect choice good for all consummated plays is for analytic purposes only and need not imply two fixed "types." In fact, the model is quite compatible with individuals' choices alternating between cooperation and defection as a consequence of whatever processes are promoting those behaviors at the time and, therefore, with the aggregate incidence of such choices also fluctuating from time to time. The model only requires that at the *point of choice* people are either cooperators or defectors.

ASSUMPTION 2. *Each person decides whether or not to play one such game with each of the others, and each player has a veto on play with each other player. A game is consummated and play completed only when two players agree to play with each other.*

The model implies no particular *sequencing* of offers to play and acceptances or rejections of those offers, such as would invite game-theoretic analyses. While it is possible that players in natural situations do sometimes adapt their cooperate–defect choices to others' invitations to play, for example, we are interested simply in the consequences of their making play–no play decisions by the heuristic in question, and cooperate–defect choices are a given in our model. For simplicity, therefore, imagine all players as making their offers to play (or refraining from such offers) at the same time.

Since t and $c > 0$ and d and $s < 0$, people will play if and only if they expect that the person they have met will cooperate. Cooperators wish to play with other cooperators because that is the only way they can make more than 0, while defectors wish to play with cooperators for the same reason. Of course, $t > c$ means that defectors make a higher return than cooperators from each successfully consummated play with a cooperator.

We model expectation formation in the particular case as subjects' taking a random sample of one from a probability distribution defining the expected incidence of cooperation in the population. Players assign some probability to meeting a cooperator and, in effect, bet in those terms that a newly met person is a cooperator. Bias one way or another in a player's expectations simply moves the probability of play and no-play decisions in the direction of the bias. It does not determine that choice: players might play even if their expectations are biased away from cooperation, and might decline to play even if their expectations are biased toward cooperation. A player's probability of being correct is, of course, defined by the actual incidence of cooperators and defectors in the population.

In natural situations, no doubt, people choose their partners by a number of criteria; but we are interested in the consequences of one that is among the simplest —projecting their own behavioral intentions onto others.

ASSUMPTION 3. *Whatever their cooperate–defect choice, people are biased toward expecting others to do the same as themselves. Cooperators judge the proportion of cooperators to be q while defectors judge it to be r, where $q >> r$.*

People are assumed to be biased in this manner whether their cooperate–defect choice is based on exclusive attention to private prisoner's dilemma incentives, on supplementary incentives (e.g., from Leviathan or decentralized normative systems), on religious or ethical principles, on the anticipation of long-term gains through reputation effects, on a "perception of themselves as individuals strongly linked to others through a shared humanity" (Monroe, Barton, and Klingemann 1989), or anything else. Nor does it matter whether the projection involves generalizing from the specifics of their own circumstances and motivations or, more simply, generalizing just from their own choice per se. All the model requires is that people figure out other people, in Mark Twain's phrase, by examining themselves.

Empirical Justification

Social Dilemma Data. In a succession of experimental studies exploring the circumstances surrounding cooperation in n-person prisoner's dilemmas, we have collected data about subjects' expectations of others' behavior.[2] One of our most consistent findings throughout these studies —a finding replicated by others' work—is that cooperators expect significantly more cooperation than do defectors. This result has been found both when payoffs are "step-level" (when contributions from a subset of *k* subjects ensure provision of a benefit to all) and when they are "symmetric" (when all contributions ensure a constant benefit to all). Table 1 reports data from a study with the former design, and Table 2 reports data from one with the latter design.

In the linear case (Table 2), we report expectations by discussion–no discussion conditions and by the target's actual cooperate–defect choice, as well as by the judge's choice. Note that cooperation is more frequent in discussion conditions than in no-discussion conditions ($p < .001$); discussion works in *some* way to increase the base of cooperation.[3]

Linkage between Behavior and Expectations. While game-theoretic formulations treat expectations about others' behavior as conditions of the actor's own behavior, an early study by Dawes, McTavish, and Shaklee (1977) reported findings consistent with expectations' being dependent on the actor's own behavior—as required by the present model. These authors solicited expectations not only from sub-

Table 1. Mean Expectations of Others' Cooperation by Judges' Cooperation–Defection Choices in a Step-Level Dilemma Experiment

	Probability of Being		
Dilemma Type	Futile	Critical	Redundant
Standard			
Cooperators	.24	.29	.46
Defectors	.55	.18	.26
Money-back guarantee			
Cooperators	.25	.29	.46
Defectors	.28	.21	.50
Enforced contribution			
Cooperators	.20	.23	.57
Defectors	.51	.25	.24

Source: Dawes et al. 1986, 1181.

Note: Seven-person groups were presented with a step-level dilemma in which a "bonus" of $10 for each member was provided for all members in the event that four or more contributed a $5 endowment toward that good. In the *standard dilemma,* contributing when three or less contributed made the contribution *futile* (a subject would lose his or her endowment); contributing when three others also contributed made the contribution *critical*; and contributing when four or more others contributed made the contribution *redundant.* In the *money-back guarantee* dilemma, subjects would get their contribution back if enough others did not contribute to produce the good. In the *enforced contribution* dilemma, subjects would be obliged to contribute if contributions from others were sufficient to produce the good. Expectations were measured by the method outlined in note 2. Relevant to this argument, there is a main effect for cooperation ($F_{1,102} = 26.36$, $p < .001$).

Table 2. Mean Expectations of Others' Cooperation by Judges' and Targets' Own Choices and by Discussion–No Discussion in a "Linear" Social Dilemma Experiment

	Condition and Judges' Choice			
	Discussion		No Discussion	
Target Choice	Cooperate	Defect	Cooperate	Defect
Cooperate	.81	.56	.63	.36
Number of cases	(104)	(104)	(68)	(74)
Defect	.57	.32	.61	.34
Number of cases	(83)	(98)	(123)	(148)

Note: The experiment on which these data are based is reported more fully as experiment 1 in Orbell, van de Kragt and Dawes 1988. This table reports expectations derived by the first method outlined in note 2. For each subject (the target) we averaged expectations held by others in the group who cooperated and defected (the judges); where one or more of the up-to-six judges in each behavioral category failed to report expectations (or misunderstood how to record their expectations), we deleted the case. Subjects had to choose between "giving" a six-dollar endowment to six "others in the experiment" (in which case each of the six benefited by $2) and "keeping" that endowment for themselves. The original design crossed (1) initial understandings about *who* would receive the benefit of a decision to give (the six others in one's own group or six matched individuals in a second group), (2) group discussion (10 minutes among the seven people in each group or no such discussion), and (3) subsequent understandings about who would receive the benefit of a decision to give. Cooperation rates varied substantially across the resultant eight conditions, but for present purposes we report expectations only by discussion.

jects who actually chose between cooperation and defection in a social dilemma game but also from subjects who only observed the game through a one-way mirror and did not make actual choices themselves. The authors reasoned that if expectations are developed independent of the individual's own choice, the standard deviation of expectations in the two cases should be similar. If, on the other hand, expectations are to some extent derived from subjects' own choices, the standard deviation of expectations for choice makers should be greater than the standard deviation for those who simply observed, because people who choose to cooperate will be led to expect cooperation, and people who choose to defect will be led to expect defection. They found that the figure for choice makers was, in fact, greater than for those who simply observed.

As Dawes (1980) has pointed out, in n-person dilemmas as distinct from two-person ones, (1) harm for defection is spread out over many players; (2) behavior is likely to be anonymous; and (3) except when all others simultaneously choose cooperation or defection, players have little or no reinforcement control over specific other individuals. No doubt these differences mean that empirical findings from social dilemma experiments should be extrapolated to two-person prisoner's dilemma games only with caution. We believe, however, that the formal similarities between the two paradigms (both involve a dominant incentive associated with a suboptimal social equilibrium) make a similar relationship between expectations and behavior in the two cases likely.

Theoretical Justification

In developing expectations about others' behavior, an individual does have information about his or her own actual or intended choice; and *in the absence of information about any other individuals' likely choices,* a reasonable assumption (at least for a Bayesean) is that his or her own choice is representative of how others will choose. Of course, if information about others were available, it would be appropriate to incorporate it into a revised estimate. But without such additional information, extrapolating from one's own choice is a reasonable thing to do. We can predict, therefore, that when the problem involves sets of strangers interacting for the first time, subjects' expectations of others' choices will be substantially biased by their own choices. (See Dawes 1988b for development of this theme in the context of the literature on "false consensus"—which literature, of course, provides further empirical support for a general tendency of people to expect others to be like themselves.)

Expected Payoffs

We can now derive expected payoffs for cooperators and defectors. Since the relative expectations are unaffected by scale changes, let $d = -1$ and, to avoid confusion, $s = -z$. A cooperator proposes play to another person with a probability q. Then, that person is

1. a cooperator with probability p, in which case the person agrees with probability q, and the payoff to the proposer is c; and
2. a defector with probability $1 - p$, in which case the person agrees with probability r, and the payoff to the proposer is $-z$.

Note that both cooperators and defectors will accept a proposal to play only if they believe the proposer is a cooperator; those who are cooperators themselves accept with an expectation of receiving c (> 0) from mutual cooperation, while those who are defectors accept with an expectation of receiving t (> 0) from unilateral defection. Hence, the expected payoff to a cooperator on encountering some other individual is $q(pqc - (1 - p)rz)$.

In contrast, a defector who meets another person proposes play with probability r. Then, that person is

1. a cooperator with probability p, in which case that person agrees to play with probability q, and the payoff to the proposer is t; and
2. a defector with probability $(1 - p)$, in which case that person agrees with probability r, and the payoff to the proposer is -1.

Hence, the expected payoff to the defector on encountering some other individual is $r(pqt - (1 - p)r)$. The difference is

$$pq(qc - rt) - (1 - p)r(qz - r), \qquad (1)$$

which can be positive. We term expression 1 the *cooperators' advantage* (which can, of course, be negative).

For example, let $t = 2$, $c = 1$, $z = 2$ (i.e., $s = -2$), $d = -1$, $p = .5$, $q = .8$, and $r = .2$. I, a cooperator, offer play with probability .8. The other individual is a cooperator *and* accepts with probability $.5 \times .8 = .4$, in which case my payoff is 1; or the other individual is a defector *and* accepts with probability $.5 \times .2 = .1$, in which case my payoff is -2. Hence, upon encountering someone, my expectation is $.8\ (.4 \times 1 - .2) = .16$. On the other hand, you, a defector, offer play with probability .2. The other individual is a cooperator *and* accepts with probability .4 (again $.5 \times .8$) and your payoff is 2; or the other is a defector *and* accepts with probability .1 ($= .5 \times .2$), in which case your payoff is -1. Hence, upon encountering someone, your expectation is $.2\ (.4 \times 2 - .1) = .14$. I have a higher expectation.[4]

Boundary Conditions

This model is restrictive. It can only produce a cooperators' advantage when d (therefore z) is negative and defectors lose some value from interacting with other defectors; clearly, natural prisoner's dilemmas in which all values are positive do exist.

Further, the model requires setting t, c, d, z, p, and r and then solving for q; and solutions *might* not be in the (0, 1) interval. Critical parameters are

1. $p/(1 - p)$, the ratio of cooperators to defectors in the population—the higher this ratio, the more opportunities cooperators have for playing with other cooperators and for capturing, therefore, cooperators' surpluses;
2. q/r, the ratio of cooperators' to defectors' optimism about cooperation in the population—the higher this ratio,
   a. the greater cooperators' gains from playing with other cooperators (pqc) relative to defectors' gains from playing with cooperators (pqt) and
   b. the smaller cooperators' losses from playing with defectors ($(1 - p)rz$) relative to defectors' losses from playing with other defectors ($(1 - p)r$); and
3. z and t—the possibility of a cooperators' advantage when cooperators and defectors play together[5] is a negative function of z and t summed.

To define the possibility of a cooperators' advantage in these terms, normalize expression 2 relative to defectors; specifically, divide by $(1 - p)r^2$:

$$\frac{p}{(1-p)}\left(\frac{q}{r}\right)\left[\left(\frac{q}{r}\right)c - t\right] - \frac{(1-p)}{(1-p)}\left[\left(\frac{q}{r}\right)z - 1\right] . \qquad (2)$$

Let $P = p/(1 - p)$, the ratio of cooperators to defectors; let $\alpha = q/r$, the ratio of cooperators' to defectors' expectation of cooperation. The cooperators' advantage is

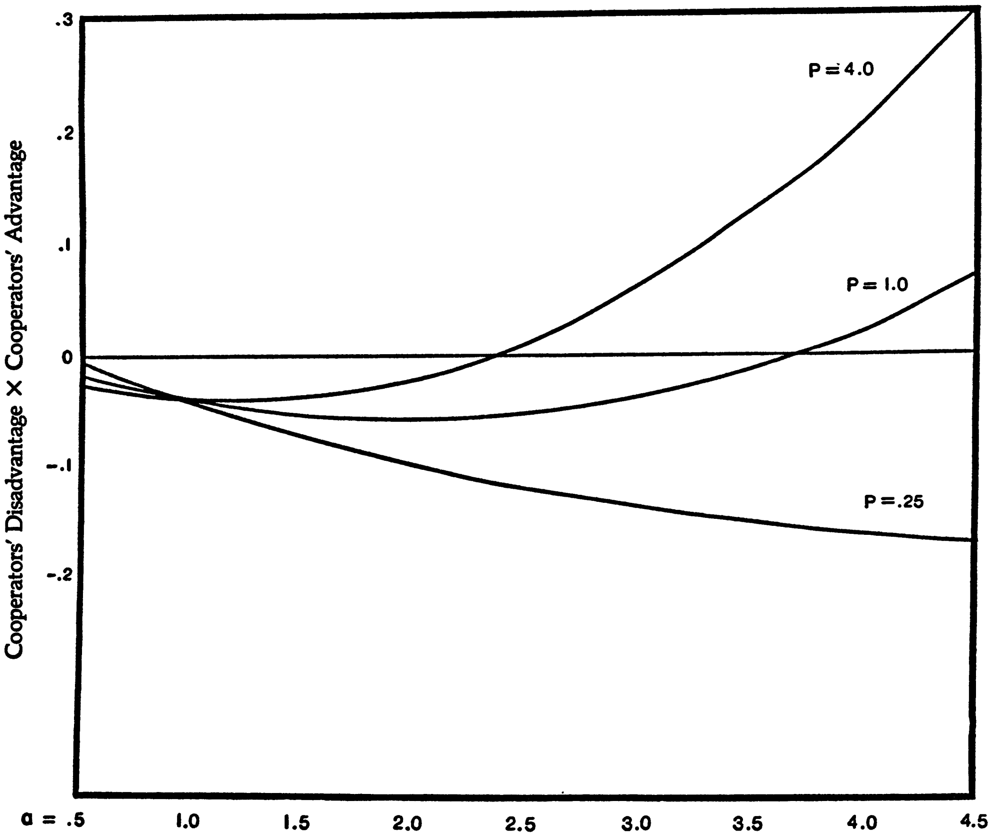
$$(2) = P\alpha\,[\alpha c - t] - [\alpha z - 1]$$
$$P c\alpha^2 - \alpha(Pt + z) + 1. \qquad (3)$$

Note that expression 3 is a linear function of P, c, t, and z. It is always a positive linear function of c and a negative one of t and z, which is consistent with the meaning of these payoffs. In contrast, it is a positive linear function of P if $c\alpha - t$ is positive, otherwise negative. (Note that by assumption $\alpha > 1$.) To define the quadratic role of α, consider the derivative of expression 3 with respect to α, which equals $2Pc\alpha - (Pt + z)$. Thus, the sum of Pt and z reduces the value of the slope in a linear way; if expression 3 is positive, increasing their sum makes a cooperators' advantage less likely, and decreasing their sum makes it more likely.

In contrast, α and P are multiplicative. Assuming that c, t, and z are set, we have an "explosion" effect: the more cooperators there are in the population, the more beneficial to them is their relative optimism, which in turn is more beneficial the more cooperators there are.

For example, Figure 1 sets t, c, d, and z as in the previous case and shows how expression 2 varies as a function of increasing α values for three values of P. The middle P value, where the population is evenly divided between cooperators and defectors, is the one from which the previous example was drawn; as can be seen, cooperators' expected payoff moves above the break-even point when α is between 3.5 and 4.0. On the other hand, with P at 4.0, cooperators have a clear advantage before α reaches 2.5. Note that the upper two functions, at least, are *U*-shaped; up to a point, cooperators are advantaged by *less* optimism about others; but beyond that point increasing

Figure 1. Cooperators' Advantage



Matrix:	1, 1	−2, 2
	2, −2	−1, −1

α = Ratio of cooperators' expectations to defectors' expectations.
P = Relative proportion cooperating in the population.

optimism pays—and pays increasingly well. Also note that all three functions would be somewhat higher if the difference between t and z were smaller. The interaction, however, is clear.

Adjustment

Colleagues to whom we have presented our model have suggested that as the size of the available sample expands through a succession of experiences, people will place less and less reliance on their own choice as a basis for predicting others' choices—that the sample will become progressively representative and that therefore beliefs of cooperators and defectors about the proportion of cooperators in the population will converge to the actual proportion. Under those circumstances, any advantage that cooperators might have had from the heuristic would evaporate.

There are reasons for expecting much less adjustment than this argument implies. First, the *need* for adjustment might not seem particularly pressing. While cooperators are involved in a higher proportion of consummated plays than defectors,[6] both experience the same high proportion of cooperative responses $= pq/(pq + (1 + (1 - p)r))$, a proportion, note, that must always be higher than p, the "veridical" proportion of cooperators.[7] Our model requires both cooperators and defectors to *believe* that they can spot cooperators; and experience in consummated plays will simply confirm their ability to discriminate.

Second, the model itself predicts differential expectations even if the same decision is repeated with outcome feedback. While cooperators and defectors both experience more cooperation than defection, both also have a high rate of not playing—either because they do not offer to play or because an offer is rejected. Moreover, the defectors have a much higher rate of not making offers, with relative proportions equal to $(1 - r)/(1 - q)$. Hence, assumptions consistent with prior belief about the propensities of the people with whom games are not consummated can leave the belief about the overall base rate of cooperation in the population minimally affected—or even unaltered—by the feedback from the consummated plays.

In the example we used above, a defector would, like a cooperator, observe an 80% cooperation rate among other players in the games actually played. But a defector offers to play with a probability of .20 and the offer is accepted with a probability of .50 ($.50 \times .80 + .50 \times .2$), leaving the defector with actual playing experience among only 10% of the population. If the defector were to believe that only .13 of this remaining 90% would have cooperated in consummated play (a figure not *very* discrepant from .20), the base rate estimate of .20 would remain unchanged.[8]

Third, the experimental evidence we cite provides no compelling reason to expect much adjustment.[9] As we have emphasized, our model refers to a single decision taken at a particular point in time. While the decisions in the experiments we cite were made in the laboratory, they were nevertheless just such decisions. The people making them came to the laboratory, no doubt, with lifelong experience in situations with prisoner's dilemma structure; but the evidence that cooperators' expectations of cooperation far exceed defectors' in these single decisions is, as we have shown, nevertheless very strong.

Group discussion in the linear dilemma case presented above (Table 2) does provide subjects with information about others' intentions (however discounted); and information is central to the issue of adjustment. In the absence of discussion, there is no contingency between targets' actual choice and judges' expectations—regardless of whether the judges were themselves cooperators or defectors. With discussion, however, the contingency is clear in both cases. More relevant to adjustment, the difference between cooperators' and defectors' expectations is present in both discussion and no-discussion conditions.

Information about others' intentions gathered from group discussion is not, of course, the same as information gathered from experience with successive plays; but it is information, nevertheless. While that information does increase subjects' accuracy, it does not reduce the *difference* between cooperators' and defectors' expectations—does not reduce, that is, the propensity of cooperators and defectors to expect others to behave as they themselves behave.[10]

It is easy to spin different hypotheses. One is that cooperators adjust more readily than defectors. Defectors encountering cooperators could well explain the event in terms of their own capacity to "spot suckers" and, accordingly, refrain from adjusting; cooperators encountering defectors, on the other hand, will have no such rationalization available to them. Early work by Kelly and Stahelski (1970) is consistent with this speculation, showing that cooperators encountering defectors are more likely to change their behavior (although not their cooperative disposition) than are defectors encountering cooperators.

Alternatively, cooperators might adjust *less* readily: Their larger number of consummated plays comprises a more reliable base for predicting to the whole population and, therefore, less reason for doubting the information it provides.

Another hypothesis says that the learning cooperators take from successive experiences is more polarized than the learning that defectors take. Cooperators encountering other cooperators will be reinforced for their generalized expectation *and* for their capacity to predict in the particular case; encountering defectors, they will be reinforced for neither. On the other hand, defectors encountering cooperators will be reinforced for their capacity to predict in the particular case but not for their generalized expectation; encountering other defectors, they will be reinforced for their generalized expectation but not for their capacity to predict.

The issue is ultimately empirical, of course. Yet adjustment becomes important only when we wish to predict behavior in a particular, ongoing social or interactive situation. Our immediate concern is to show only that the expectations heuristic is sufficient—under appropriate circumstances—to produce a cooperators' advantage and to provide a basis for evolutionary expansion of that advantage. The extent to which it is used in natural situations and the circumstances under which it is rejected are matters for further development.

Discussion

The expectations heuristic is not sufficient to produce a cooperators' advantage in the absence of other mechanisms supporting it. If there are no other cooperators in the population, to illustrate with the extreme case, a single cooperator will lose absolutely and comparatively from every play he or she consummates.[11]

Yet this is not a particular weakness of the model. Despite some extravagant claims (e.g., for the necessity of "incentive compatibililty"), it is unlikely that so basic a process as cooperation is explained by a single mechanism; and interactions (e.g., with Hobbesian coercion or evolved norms) are to be expected. Maruyama (1963) has shown that an "initial kick" (which might never be known) can trigger a deviation-amplifying mechanism that propagates and enhances that initial change. The heuristic we have discussed could act as just such a deviation amplifier in the case of cooperative behavior.

We have assumed that all individuals can interact with all others in the population; but relaxing that to allow a greater probability of interacting among adjacent pairs could give cooperators a higher probability of interacting with other cooperators in certain locations even if there were relatively few in the wider set. In evolutionary terms, the cooperators' surplus thereby produced would increase the relative strength of cooperation in such locations in the next generation, with strength feeding on strength through successive generations.[12]

While we assume that expectations (of cooperators and defectors alike) have only random predictive capacity, we do not deny that translucency might have evolved. As Gauthier (1986) suggests, there is a continuum between transparency, translucency, and opaqueness. Transparency, of course, would ensure a cooperators' advantage; if everyone could recognize everyone else's intentions, no defector would enjoy a consummated relationship. If arguments such as Frank's (1988) on the role of emotions in cue giving, and Cosmides' (1989) on special purpose cognitive mechanisms for detecting cheating are correct, people might be to *some* extent translucent—certainly more than is implied by our conservative assumption of a random hit rate.

It might be suggested that our heuristic identifies the lower bound of cooperators' advantage that is available from selective play. Perhaps, but in the absence of group discussion, we have little evidence of the capacity to predict others' behavior, as required by other models of cooperators' advantage from selective play. In Table 2 we have shown that while judges were reasonably successful in discriminating between cooperating and defecting targets after group discussion, they were not at all successful without it. Similarly, Dawes, McTavish, and Shaklee (1977) found significant accuracy in communication conditions (which, they comment, is "not exactly surprising") but otherwise found little more than random accuracy in one condition, and a slight effect in the opposite direction in another.[13] Further empirical investigation might show that the heuristic discussed here is the *only* basis for cooperators' advantage via selective play when there is no opportunity for sustained, personalized interaction.

Our model incorporates the everyday observation that people who are themselves untrustworthy in joint ventures tend not to trust others and are less willing than optimistic cooperators to seek out or agree to such ventures. Successful enterprise requires that offers for cooperative interaction be accepted by others who themselves have a stake in that interaction; and our model shows how entrepreneurs who intend honest cooperation can, in fact, end up ahead of those who intend exploitation of potential partners. While the latter *are* exploited less often than the former, they also miss opportunities for personal gain that depend on mutual cooperation.

Perhaps paradoxically, the cooperators' advantage we have identified is not available through calculated, strategic play.[14] It depends, simply, on optimism about others' behavior; and such optimism cannot be invented in the interest of promoting private advantage.

Notes

We thank Deborah Baumgold, Linnda Caporael, Robert Clemen, John Dryzek, Deborah Frisch, David Gauthier, David Goetze, George Loewenstein, Gerry Mackie, Robert Mauro, Tom Palfrey, Myron Rothbart, Perigrine Schwartz-Shea, Viktor Vanberg, Rick Wilson, and Tom Youderian for helpful comments.

1. That is to say, t = *temptation* resulting from unilateral defection; c = *cooperate* resulting from mutual cooperation; d = *defect* resulting from mutual defection; and s = *sucker* resulting from unilateral cooperation. Of course, the values specified here represent only a subset of possible prisoner's dilemma games. In some cases, all values will be

positive (with players making gains of varying magnitudes by comparison with the status quo); and in some cases, all values will be negative (with players making losses of varying magnitudes).

2. Although the specifics have varied somewhat between experiments, we have two basic techniques for eliciting such expectations. First, we ask each subject to assign a score between 100 and 0 to every other subject, the particular value indicating "how probable" the judge sees the target's cooperation (*X*-choice); a score of 50, of course, indicates that the judge cannot decide. Second, we ask each subject to assign a score between 100 and 0 to each of the possible outcomes at the group level (e.g., 0 choose *X*; 1 chooses *X* . . . , *n* choose *X*) with the magnitude of the score indicating "how probable" the judge believes that outcome to be and the scores for all $n + 1$ outcomes summing to 100.

3. Orbell, van de Kragt, and Dawes (1988) identify discussion-induced group solidarity and promise making under a rule requiring everyone in the group to have promised as possible explanations of discussion's much replicated effect; but it remains to separate the roles of these two variables or to identify possible interaction between them.

4. Transaction costs might modify this outcome. The cooperators' advantage depends on cooperators' having more plays (with cooperators) than do defectors, and it could be reduced by per-play costs of interacting. Alternatively, the issue of transaction costs could be dealt with by subtracting them from the payoffs in the matrix. However, transaction costs seem likely only to reduce the possibility of a cooperators' advantage, not eliminate it.

5. An error, notice, for cooperators but a success for defectors.

6. Because of this, the set of consummated interactions will be a poor base from which to assess the incidence of cooperators and defectors in the general population. Cooperators' optimism about others means that cooperators will be overrepresented in that set, just as defectors' pessimism means that defectors will be underrepresented. Observing that, say, 75% of players in consummated interactions cooperate, an optimist might conclude that 75% of the population is cooperative. The true figure would be substantially lower.

7. Note, in addition, that if it is also higher than *q* (hence *r*), both cooperators and defectors are reinforced in their belief that they are good at spotting cooperators (i.e., are "above average judges of character"—like everybody else).

8. Of course, the implications that cooperators and defectors draw from others' rejecting their offers of play remains an empirical question. Given everyday observation and empirical data about "projections" from trustworthy and untrustworthy people, it seems possible that defectors believe they are turned down because cooperators "see through" their defecting intentions, while cooperators interpret such rejection as reflecting a neurotic lack of trust in their own cooperative intentions.

9. The extensive literature on iterated two-person prisoner's dilemmas is relevant to adjustment when subjects play a series with the same individual; but that is an issue distinct from adjustment when subjects play single games with a succession of different individuals from some population. Hoch and Loewenstein (1989) have provided evidence that at least one other decision heuristic, "hindsight bias," is resistant to outcome feedback. Such feedback can reduce the effect of that bias, but not always; in some cases, the heuristic is strong enough to overwhelm the information contained in the feedback (see also Sherman, Presson, and Chassin 1984).

10. This despite the fact that discussion substantially increases the incidence of cooperation—meaning that (in Figure 1's terms) α does not vary with *P*.

11. Generally, with prisoner's dilemma payoff parameters constant, a cooperators' advantage requires that the difference between *q* and *r* become greater as the proportion of cooperators in the population becomes smaller. Empirical investigation might reveal a threshold of cooperation that must be reached before the mechanism we have identified can begin to operate to cooperators' advantage.

12. Comparable arguments are made by Axelrod (1984) and by Vanberg and Buchanan (1988).

13. Frank (1988) designed experiments explicitly to assess subjects' capacity to predict cooperation after a thirty-minute "get acquainted" period during which they were free to discuss any subject they wished. Consistent with the data in Table 2 and with Dawes, McTavish, and Shaklee (1977), he found considerable predictive success. He did not, however, test such predictive accuracy in the absence of discussion.

14. As pointed out to us by Thomas Palfrey (personal communication 23 May 1990), strategic behavior that allows choice of the highest expected value strategy (i.e., defection) after assessing others' cooperative propensity on the basis of one's own effectively erases the cooperators' advantage.

References

Alexander, Richard. 1979. *Darwinism and Human Affairs.* Seattle: University of Washington Press.

Alexander, Richard. 1987. *The Biology of Moral Systems.* New York: De Gruyter.

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Bell, David, Howard Raiffa, and Amos Tversky, eds. 1988. *Decision Making: Descriptive, Normative, and Prescriptive Interactions.* Cambridge: Cambridge University Press.

Boyd, R., and P. Richerson. 1988. "The Evolution of Reciprocity in Sizable Groups." *Journal of Theoretical Biology* 132:337–56.

Brennan, Geoffrey, and James Buchanan. 1985. *The Reason of Rules.* Cambridge: Cambridge University Press.

Buchanan, James. 1965. "Ethical Rules, Expected Values, and Large Numbers." *Ethics* 76:1–13.

Buchanan, James. 1972. *The Bases for Collective Action.* New York: General Learning.

Buchanan, James. 1977. *Freedom in Constitutional Contract.* College Station: Texas A&M University Press.

Buchanan, James. 1978. "Markets, States, and the Extent of Morals." *American Economic Review* 68:364–68.

Campbell, D. 1975. "On the Conflicts between Biological and Social Evolution and between Psychology and Moral Tradition." *American Psychologist* 30:1103–26.

Cosmides, L. 1989. "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task." *Cognition* 31:187–276.

Dawes, Robyn. 1980. "Social Dilemmas." *Annual Review of Psychology* 31:169–93.

Dawes, Robyn. 1988a. *Rational Choice in an Uncertain World.* San Diego: Harcourt, Brace, Jovanovich.

Dawes, Robyn. 1988b. "Statistical Criteria for Establishing a Truly False Consensus Effect." *Journal of Experimental Social Psychology* 25: 1–17.

Dawes, Robyn, J. McTavish, and H. Shaklee. 1977. "Behavior, Communication, and Assumptions about Other Peoples' Behavior in a Commons Dilemma Situation." *Journal of Personality and Social Psychology* 35:1–11.

Dawes, Robyn, J. Orbell, R. Simmons, and A. van de Kragt. 1986. "Organizing Groups for Collective Action." *American Political Science Review* 80:1171–85.

Frank, Robert. 1988. *Passions within Reason.* New York: Norton.

Gauthier, David. 1986. *Morals by Agreement.* Oxford: Clarendon.

Hobbes, Thomas. 1947. *Leviathan.* London: J. M. Dent.

Hoch, S., and G. Loewenstein. 1989. "Outcome Feedback: Hindsight and Information." *Journal of Experimental Psychology; Learning, Memory, and Cognition* 15:605–19.

Hogarth, Robin, and Melvin Reder, eds. 1987. *Rational Choice: The Contrast between Economics and Psychology.* Chicago: University of Chicago Press.

Kelly, H., and A. Stahelski. 1970. "Social Interaction Basis of Cooperators' and Competitors' Beliefs about Others." *Journal of Personality and Social Psychology* 16:66–91.

Maruyama, M. 1963. "The Second Cybernetics: Deviation Amplifying Mutual Causal Processes." *American Scientist* 51(2):164–79.

Masters, Roger. 1989. *The Nature of Politics.* New Haven: Yale University Press.

Monroe, K., M. Barton, and U. Klingemann. 1988. "Altruism and the Theory of Rational Action: An Analysis of Rescuers of Jews in Nazi-Europe." Presented at the annual meeting of the Western Political Science Association, Salt Lake City.

Orbell, J., A. van de Kragt, and R. Dawes. 1988. "Explaining Discussion-induced Cooperation." *Journl of Personality and Social Psychology* 54: 811–19.

Parsons, Talcott. 1937. *The Structure of Social Action.* New York: McGraw-Hill.

Richerson, P., and R. Boyd. 1988. "The Role of Evolved Predispositions in Cultural Evolution; or, Human Sociobiology meets Pascal's Wager." *Ethology and Sociobiology* 10:1–25.

Sherman, S., V. Presson, and L. Chassin. 1984. "Mechanisms Underlying the False Consensus Effect: The Special Role of Threats to the Self." *Personality and Social Psychology Bulletin* 10: 127–38.

Singer, Peter. 1983. *The Expanding Circle.* Oxford: Oxford University Press.

Taylor, Michael. 1976. *Anarchy and Cooperation.* New York: Wiley & Sons.

Tetlock, P. 1979. "Accountability: The Neglected Social Context of Judgment and Choice." *Research in Organizational Behavior* 7:297–332.

Trivers, R. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46: 35–57.

Vanberg, V., and J. Buchanan. 1988. "Rational Choice and Moral Order." *Analyse und Kritik* 10:138–60.

Wilson, David. 1980. *The Natural Selection of-Populations and Communities.* Menlo Park, CA: Benjamin/Cummings.

Wilson, Edward. 1975. *Sociobiology.* Cambridge: Harvard University Press.

John M. Orbell is Professor of Political Science, University of Oregon, Eugene, OR 97403.

Robyn M. Dawes is Professor of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213.